JISC Shared Infrastructure Services Workshop
26 June 2007, Brettenham House, London

High-Level Thesaurus (HILT) project, phase IV
Centre for Digital Library Research & EDINA

George Macgregor
Centre for Digital Library Research
University of Strathclyde

The problem the service is addressing

As it becomes increasingly difficult for users to satisfy their information needs due to the rapid expansion of the Web, it is also becoming progressively impractical for users to consult a wide range of sources to satisfy an information query. Consequently, it is of growing importance that users are able to search multiple distributed heterogeneous digital repositories simultaneously. With such a wide variety of resources available, however, the feasibility of achieving interoperability between them is gradually diminishing. Not only do services employ different technical standards, indexing practices, search facilities and algorithms, but also the basic language on which retrieval systems are founded differs widely. It is no longer sufficient for users to make decisions on whether to use keyword or phrase searching, employ Boolean operators, or try their luck with truncation, they must also now give consideration to the terminology they use.

Problems relating to disparate terminology use have been an impediment to information retrieval for many years, but the growth of Web, associated heterogeneous digital repositories, and the need for distributed searching within multi-scheme information environments (such as the JISC IE) has recently drawn the issue into sharp focus. The HILT project, which is now in phase IV, aims to research, investigate and develop pilot solutions for problems pertaining to cross-searching multi-subject scheme information environments, as well as providing a variety of other terminological searching aids.

The way in which HILT is addressing the problem

HILT is currently addressing the above noted interoperability issues by providing access to an M2M terminology server. This server provides terminology mappings via a Dewey Decimal Classification (DDC) switching spine to facilitate interoperability between disparate subject schemes. This server also provides access to detailed terminological datasets, thus allowing local services to incorporate improved terminology-based searching and browsing tools (e.g. hierarchical browsing structures, terminology-based query expansion techniques, etc.). The current demonstrator of HILT is an M2M implementation, based on Web service protocols. HILT offers access via the (SOAP-based) SRW protocol and uses W3C SKOS to structure terminological data sent to clients. This demonstrator (developed in phase III) forms the basis for the design of the initial entry-level service to be built in phase IV.

The transition to service phase of HILT has just begun (late April 2007). This phase of HILT will ultimately allow an initial entry-level service to be built, tested for user requirements and retrieval effectiveness, refined in line with the findings, and extended to permit the use of a range of distributed terminology services for interoperability. This entails the creation of an initial entry-level terminologies and subject interoperability service, comprising a freely available package consisting of an SRW client with illustrative user interface routines (which could be customised by local JISC information services) to exploit HILT facilities,

terminologies, and terminology mappings. The initial entry-level service will also provide a terminologies database with high-level and selected in-depth mapping, a SOAP-based HILT requests and responses handler based around a series of search and retrieve functions identified in HILT phase III, an SRW server providing a standard interface to the SOAP requests and responses handler, and client use of IESR and the HILT database of terminologies and mappings to identify collections appropriate to a user's subject request.

HILT phase IV will also examine a variety of other issues pertaining to client interface and retrieval requirements (including HILT system evaluation), dissemination and survey activity, a report on research into various selected issues of relevance to the provision of an effective future entry-level service or its further refinement, and the development of future proposals (including cost and maintenance estimates, etc.)

Benefits to users

The main benefits to users include the following:

- Improved subject interoperability between services, thereby improving the ability of users to conveniently and accurately search/browse distributed heterogeneous collections/services simultaneously within the JISC IE (and beyond!). This is expected to be the primary benefit to users; expanding the corpus of resources capable of being searched simultaneously, and increasing the precision of the results returned by accurately matching user terms to local indexes. It is also worth noting that a general benefit of HILT is that it is deployable in an M2M context; user understanding or interaction with HILT is unnecessary as client interaction with HILT occurs 'in the background'.
- There is potential for improved searching and browsing of locally indexed content for users if local services incorporate some of the generic HILT server functions. This might entail the provision broader / narrower terms, related terms, etc. or other forms of terminology-based interactive query expansion, or dynamic hierarchical browse structures, graph visualisations, and so forth. The terminological data requested can be used by clients as they wish.
- There are opportunities for clients to use IESR and the HILT database of terminologies and mappings to identify collections appropriate to users' subject requests, determine the subject schemes they use, and provide subject interoperability by offering subject access via scheme hierarchies entered at a point appropriate to users' subject interest (so-called 'directed entry').

What is currently available for implementation?

Nothing is available for implementation from HILT phase IV as this phase of the project has only just begun; however, some of the demonstrators from HILT phase III are available for experimentation:

- HILT SOAP client demonstrator is available (http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/hiltsoapclient.php). Details on M2M connection and functions are available in the associated explain file (http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/hiltsoapclient.php?request=get_explain). Contact e.mcculloch@strath.ac.uk or anu.joseph@cis.strath.ac.uk if you experience connection difficulties.

Some implementations demonstrating the use of HILT are also available. See for example:

- HILT SRW client demonstration (HILT II emulation using SRW and SKOS Core): http://hilt3.cdlr.strath.ac.uk/
- HILT SRW client 'scheme specific browse' demonstrator (using SKOS Core): http://hilt3.cdlr.strath.ac.uk/hilt_srw.cgi

George Macgregor,
Centre for Digital Library Research (CDLR),
Department of Computer & Information Sciences,
University of Strathclyde, Livingstone Tower,
26 Richmond Street, Glasgow, UK, G1 1XH
tel: +44 (0)141 548 4752
fax: +44 (0)141 548 4523
web: http://cdlr.strath.ac.uk/
FOAF: http://cdlr.strath.ac.uk/foaf/george.rdf